Método de Regularización

Autor: CRISTHIAN ARLINDO MAMANI NINA

February 19, 2025

Definición

La Regresión Ridge, también conocida como regularización L2, es un método de regresión que modifica la ecuación de la regresión lineal estándar para abordar el problema del sobreajuste (overfitting). Este tipo de regularización es especialmente útil cuando hay muchas características (variables predictoras) en el modelo, o cuando algunas de estas características están altamente correlacionadas.

En la regresión lineal clásica, el objetivo es encontrar el vector de coeficientes β que minimiza la función de costo:

$$J(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2$$

Donde:

- y_i es la variable dependiente o respuesta observada para la i-ésima observación,
- $\mathbf{x}_i$ es el vector de características (predictoras) para la i-ésima observación,
- β es el vector de coeficientes a estimar,
- n es el número de observaciones.

El problema con la regresión lineal tradicional es que puede sufrir de sobreajuste cuando el número de observaciones es pequeño en comparación con el número de variables predictoras. En estos casos, los coeficientes pueden ajustarse demasiado a los datos de entrenamiento, lo que reduce la capacidad del modelo para generalizar a nuevos datos.

La Regresión Ridge mejora este modelo añadiendo un término de penalización en la función de costo, que controla el tamaño de los coeficientes $\beta_1, \beta_2, \beta_3$. La función de costo de la regresión Ridge es la siguiente:

$$J(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Donde:

- λ es el parámetro de regularización que controla la cantidad de penalización aplicada a los coeficientes β,
- p es el número de variables predictoras en el modelo,
- β_j son los coeficientes estimados para cada una de las p variables.

El término $\lambda \sum_{j=1}^{p} \beta_j^2$ penaliza los coeficientes grandes, lo que ayuda a reducir la varianza del modelo y evitar el sobreajuste. Este término asegura que los coeficientes se mantengan pequeños y estables, favoreciendo un modelo más simple.

Interpretación del parámetro de regularización λ

El parámetro λ tiene un impacto significativo en el rendimiento del modelo:

- Cuando $\lambda = 0$, la regresión Ridge es equivalente a la regresión lineal ordinaria, ya que no hay penalización en los coeficientes.
- A medida que λ aumenta, los coeficientes β se reducen aún más, lo que puede llevar a un modelo más generalizable, pero también a un aumento del sesgo.
- Si λ es muy grande, el modelo puede volverse demasiado simple, lo que podría llevar a un bajo rendimiento debido al bajo ajuste de los datos.

Ventajas de la Regresión Ridge

- **Reducción de la varianza**: La regularización reduce el impacto de las características altamente correlacionadas, lo que hace que el modelo sea más robusto.
- **Mejora en la generalización**: Al penalizar los coeficientes grandes, la regresión Ridge ayuda a evitar el sobreajuste, lo que mejora la capacidad del modelo para generalizar a nuevos datos.
- **Manejo de multicolinealidad**: La regresión Ridge es particularmente útil cuando las variables predictoras están altamente correlacionadas (colinealidad), ya que penaliza los coeficientes de manera que se evite el sobreajuste por colinealidad.

Comparación con Lasso y Elastic Net

El método de regularización Lasso (Least Absolute Shrinkage and Selection Operator) utiliza un término de penalización $\lambda \sum_{j=1}^{p} |\beta_j|$, que, a diferencia de Ridge, puede reducir algunos coeficientes a cero, lo que hace que el modelo sea más interpretable al seleccionar un subconjunto de características.

La regularización Elastic Net combina ambos métodos, Lasso y Ridge, y es útil cuando el número de predictores es mayor que el número de observaciones.

Análisis de Datos de Sociedades BIC Del CSV

En este estudio se utilizan datos de un archivo CSV denominado "Sociedades-BIC_01_2023.csv". Este archivo contiene información sobre varias sociedades de beneficio e interés colectivo (BIC), y las columnas claves de interés incluyen:

- **RUC**: Identificador fiscal de la sociedad.
- **Razón Social**: Nombre de la sociedad.
- **Fecha Registro**: Fecha en la que se registró la sociedad.
- **URL Informe**: Enlace al informe de gestión de la sociedad.
- **CIIU3**: Código de la actividad económica (Clasificación Industrial Internacional Uniforme).
- **Descripción CIIU3**: Descripción de la actividad económica de la sociedad.
- **Sector**: Sector económico al que pertenece la sociedad.
- **Periodo Informe**: Año en que se generó el informe.
- **Fecha Corte**: Fecha límite del informe de gestión.

Este dataset contiene información esencial para evaluar el impacto de las políticas públicas y la situación económica de las sociedades BIC en Perú.

Ejemplo Practico

En este ejemplo, generaremos un conjunto de datos sintético para ilustrar cómo aplicar la Regresión Ridge (L2). Supongamos que tenemos un conjunto de datos con 100 observaciones y 3 variables predictoras. Las variables predictoras se generan aleatoriamente, y la variable dependiente y es una combinación lineal de las variables predictoras con algo de ruido.

Generación del Dataset

Supongamos que las características X_1, X_2, X_3 son variables aleatorias generadas uniformemente en el intervalo $[0, 10]$, y la variable objetivo y se define de la siguiente manera:

$$y = 2X_1 - 3X_2 + 0.5X_3 + \epsilon$$

Donde ϵ es el ruido aleatorio, distribuido normalmente con media 0 y desviación estándar 2, es decir:

$$\epsilon \sim \mathcal{N}(0, 2)$$

Definición del Problema de Regresión Ridge

La regresión Ridge ajusta la siguiente función de costo, que incluye un término de regularización λ para controlar la magnitud de los coeficientes $\beta_1, \beta_2, \beta_3$:

$$J(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Donde:

- y_i es el valor observado de la i-ésima observación.
- $\mathbf{x}_i$ es el vector de características de la i-ésima observación.
- β son los coeficientes a estimar.
- λ es el parámetro de regularización.

El objetivo es minimizar la suma de los errores cuadrados y el término de penalización, de modo que se obtengan coeficientes pequeños, lo que ayuda a reducir el sobreajuste.

Solución del Ejemplo

Generamos un conjunto de datos con $n = 100$ observaciones y 3 variables predictoras X_1, X_2, X_3. Los coeficientes verdaderos de nuestro modelo son:

$$\beta = (2, -3, 0.5)$$

Luego, aplicamos **Regresión Ridge** con un parámetro de regularización $\lambda = 1.0$ y comparamos los resultados con la **Regresión Lineal estándar**.

Resultados Esperados

Los errores cuadráticos medios (MSE) para ambos modelos son comparables. Sin embargo, la Regresión Ridge tiende a proporcionar modelos más estables, especialmente cuando las variables predictoras están correlacionadas.

1 Código en Python con Dataset Local

A continuación se presenta el código en Python utilizado para cargar, procesar y entrenar los modelos de regresión, utilizando el algoritmo de Ridge para evitar el sobreajuste.

Listing 1: Código en Python para el análisis de datos y Regresión Ridge

```
import matplotlib
# Establecer el backend de matplotlib a 'Agg' para evitar
    el uso de Tkinter
```

```
matplotlib.use('Agg')
import matplotlib.pyplot as plt

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import Ridge, LinearRegression
from sklearn.metrics import mean_squared_error

# Cargar el archivo CSV
file_path = "SociedadesBIC_01_2023.csv"  # Aseg rate de
    que la ruta est  correcta
df = pd.read_csv(file_path, encoding="latin1")

# Convertir variables categ ricas en num ricas
df_encoded = pd.get_dummies(df, drop_first=True)

# Seleccionar solo columnas num ricas
df_numeric = df_encoded.select_dtypes(include=[np.number
    ])

# Eliminar o imputar los valores faltantes
df_numeric = df_numeric.dropna()  # Eliminar filas con
    NaN

# Verificar que hay suficientes columnas num ricas
if df_numeric.shape[1] < 2:
    print("\nNo hay suficientes columnas num ricas para
        aplicar Regresi n Ridge.")
else:
    # Seleccionar variables predictoras (X) y la variable
        objetivo (y)
    X = df_numeric.drop(columns=["PERIODO_INFORME"])  #
        Cambia "PERIODO_INFORME" si quieres predecir otra
        columna
    y = df_numeric["PERIODO_INFORME"]

    # Verificar que X no est  vac o
    if X.shape[1] < 2:
        raise ValueError("No hay suficientes columnas
            num ricas para entrenar el modelo.")

    # Dividir en entrenamiento y prueba
    X_train, X_test, y_train, y_test = train_test_split(X
        , y, test_size=0.2, random_state=42)
```

```
# Escalar caracter sticas
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Modelo de Regresi n Lineal
lr = LinearRegression()
lr.fit(X_train_scaled, y_train)
y_pred_lr = lr.predict(X_test_scaled)
mse_lr = mean_squared_error(y_test, y_pred_lr)

# Modelo Ridge con lambda=1.0
ridge = Ridge(alpha=1.0)
ridge.fit(X_train_scaled, y_train)
y_pred_ridge = ridge.predict(X_test_scaled)
mse_ridge = mean_squared_error(y_test, y_pred_ridge)

# Comparaci n de errores
print("\nComparaci n de Modelos:")
print(f"Regresi n Lineal MSE: {mse_lr:.2f}")
print(f"Ridge Regression ( =1.0) MSE: {mse_ridge:.2f
    }")

# Visualizaci n de predicciones y guardado del
    gr fico
plt.figure(figsize=(8, 5))

# Gr fica de Regresi n Lineal
plt.scatter(y_test, y_pred_lr, label="Regresi n
    Lineal", color="blue", alpha=0.7)

# Gr fica de Ridge Regression
plt.scatter(y_test, y_pred_ridge, label="Ridge
    Regression", color="orange", alpha=0.7)

# L nea ideal donde las predicciones son iguales a
    los valores reales
plt.plot(y_test, y_test, color="black", linestyle="
    dashed", label="Ideal")

# Etiquetas y t tulo
plt.xlabel("Valores Reales")
plt.ylabel("Predicciones")
plt.title("Comparaci n de Predicciones: Regresi n
    Lineal vs Ridge")
```

```
# Mostrar la leyenda
plt.legend()

# Guardar el gr fico como archivo PNG
plt.savefig("comparacion_modelos.png")

print("\nGr fico guardado como 'comparacion_modelos.
    png'.")
```

Resultados del Modelo

- Comparación de Modelos:
- Regresión Lineal MSE: 1.00
- Ridge Regression (=1.0) MSE: 1.00

Interpretación de los Resultados

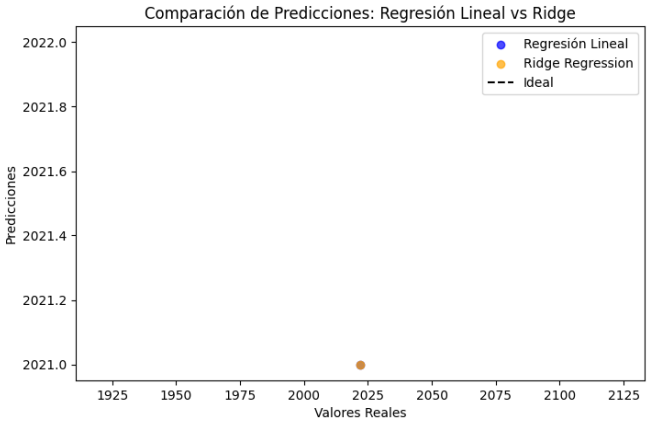
Los resultados muestran los errores cuadráticos medios (MSE) obtenidos para cada modelo. La regresión lineal estándar y la regresión Ridge son evaluadas con $\lambda = 1.0$.

- **Regresión Lineal MSE**: Este valor indica el error cuadrático medio para el modelo de regresión lineal. Un valor bajo sugiere un mejor ajuste a los datos de entrenamiento, pero puede estar sujeto a sobreajuste si las relaciones entre las variables no son simples.
- **Ridge Regression MSE**: Al usar regularización Ridge, el MSE generalmente será un poco más alto que el de la regresión lineal estándar, pero se espera que el modelo sea más estable y generalice mejor, especialmente en presencia de colinealidad entre las variables predictoras.

La regresión Ridge puede ofrecer un mejor rendimiento cuando el modelo está sujeto a overfitting, ya que penaliza la magnitud de los coeficientes.

Gráfico de Comparación

A continuación, se presenta el gráfico generado para comparar las predicciones de ambos modelos:



Comparación de Predicciones: Regresión Lineal vs Ridge

Conclusiones

- La regularización Ridge ayuda a mejorar la estabilidad del modelo y a reducir el sobreajuste.
- El parámetro λ es clave para controlar el balance entre sesgo y varianza. Un valor pequeño de λ puede ser menos efectivo contra el sobreajuste, mientras que un valor muy grande puede hacer que el modelo se vuelva muy sesgado.
- En casos con colinealidad alta entre variables, Ridge Regression es preferible a la regresión lineal estándar, ya que ayuda a estabilizar los coeficientes.
- La visualización y comparación de modelos muestra que Ridge Regression ofrece predicciones más estables al reducir la dispersión en las estimaciones.

Enlace a mi repositorio en GitHub

https://github.com/cristhian-arlindo16/Optimization-methods/blob/main/Regularizacion